W620–W623 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh457

FAN: fingerprint analysis of nucleotide sequences

Neil Maudling* **and Teresa K. Attwood**

School of Biological Sciences and Department of Computer Science, University of Manchester, Manchester M13 9PT, UK

Received February 12, 2004; Revised April 18, 2004; Accepted April 26, 2004

ABSTRACT

FAN is a server for fingerprint analysis of nucleotide sequences. The server performs a search of submitted nucleotide sequences against the PRINTS database. Searches are performed directly against fingerprints using a codon position specific score matrix (PSSM) approach. The advantages of this approach are increased specificity for coding sequence (CDS) over non-CDS, and increased tolerance to base-substituting and frameshifting sequence errors. Furthermore, there is no need for prior translation of the nucleotide sequences. A web-based interface to the software is available at http://bioinf.man.ac.uk/cgi-bin/neil/ntfront.pl.

INTRODUCTION

Nucleotide sequences are a rich source of biological data. Currently, there are 33 000 million nucleotides deposited in the GenBank database, the majority of these data coming from the various completed and ongoing genome and EST (expressed sequence tag) sequencing projects. The elucidation of novel genes in these sequences is vital to biological research, not just for the gene sequences themselves, but because knowledge of novel gene locations can disclose a wealth of other information; for instance, intron/exon boundary sequences, transcription initiation sites and gene structures can all be investigated. However, prediction of the location and structure of novel genes is a difficult problem. This is illustrated by the fact that since the draft human genome sequence was released, estimates of the number of human genes have varied widely, from approximately 28 000 to in excess of 100 000 (1). There are two major strategies for locating novel genes: use of sequence similarity to infer homology and *ab initio* gene prediction. The former usually involves translation-based methods, where six-frame translations are compared with protein sequences/alignments and homology is inferred where the sequence similarity is relatively high (typically ⩾20%). The latter can find genes that have no known homologues, but prediction accuracy is generally poor (2). With either method, gene identifications may be erroneous, or genes may be missed entirely.

Protein family databases are based on multiple alignments of known families or domains, and can often allow inference of homology where pairwise search methods fail, because of the greater evolutionary information afforded by alignments. PRINTS (3) is one such family database, which houses multiple conserved motifs in the form of family fingerprints. A motif is a short, ungapped conserved region excised from an alignment of related sequences. Fingerprints are ordered series of motifs, the order being an important discriminatory component between true and false family members. Currently, like most family databases, PRINTS lacks an in-house dedicated nucleotide sequence search tool. However, the Blocks server (4) does provide a nucleotide search of PRINTS via a six-frame translation.

Nucleotide triplet frequencies differ markedly between coding sequence (CDS) and non-CDS, and thus approaches based on codon-usage statistics have been frequently used in gene prediction tools. FAN is a new tool for fingerprint analysis of nucleotide sequences, using an approach based upon codon position specific score matrices (PSSMs). The use of cPSSMs that combine codon-usage statistics with traditional amino acid PSSMs results in matrices capable of direct comparison with DNA. In addition, it confers some specificity for CDS over non-CDS and allows the adoption of pseudocount schemes capable of tolerating sequence errors.

CODON POSITION SPECIFIC SCORE MATRICES

PSSMs and regular expressions (regexes) are standard ways of encoding motifs into descriptors capable of being used to search sequences. PSSMs offer a number of advantages over regexes: they quantify the strength of a match (i.e. they provide a score) and they allow calculation of P-values, which in turn improves the sensitivity of sorted hit lists. Ordinarily, to use a PSSM to search nucleotide sequences, six-frame translation of the DNA is required. This process, whilst increasing the search space 6-fold, results in a loss of potentially useful nucleotide compositional information. For instance, some *ab initio* gene prediction tools utilize the different triplet and 6mer frequencies in the open reading frame

*To whom correspondence should be addressed. Tel: +44 161 275 5980; Fax: +44 161 275 5082; Email: neil@bioinf.man.ac.uk

(ORF) to distinguish CDS from non-CDS (5). Such compositional discrimination can be introduced into a PSSM by combining codon-usage statistics with an amino acid PSSM to produce a new codon- PSSM (cPSSM).

PSSM searches are usually performed using a sliding window approach, where the matrix is compared with a subsequence the width of the motif. Residue scores are determined from the matrix and summed (if a log-odds matrix) or multiplied (if a raw probability matrix). For an amino acid PSSM, the probability that an individual amino acid matches the residue found in the sequence, $P(A)_i$, can be used as the score, and this can be estimated very simply from the relative frequency of the amino acid in the motif. However, for a codon matrix, the probability of obtaining the codon observed, $P(C)_i$, is required.

Obtaining $P(C)_i$

Codon-usage statistics are usually quoted as frequencies per 1000 codons $f(C)_i$, and can be used to estimate the probability of obtaining a particular codon $(C)_i$ at a sequence position given the corresponding amino acid, A_i. This estimate is obtained using Equation 1:

$$P(C_i \mid A_i) = \frac{f(C)_i}{\sum_{j=1}^{n} f(C)_j}, \tag{1}$$

where n is the number of codons that encode amino acid A_i.

However, a codon matrix score, $P(C)_i$ is required. Substituting Equation 1 into Bayes' theorem (6), gives Equation 2:

$$P(C)_i = \frac{f(C)_i \times P(A)_i}{\sum_{j=1}^{n} f(C)_j}. \tag{2}$$

Thus, a 64-codon matrix can be created by using Equation 2 for all codons at each column in the original PSSM. The cPSSM can be used much like its amino acid counterpart, except that the matrix is moved across the sequence a single base, rather than a whole codon, at a time and, within each window, reverse complement codons are also scored in order to score all possible reading frames. The cPSSM approach has several advantages:

(i) Translation of the original sequence is no longer required.
(ii) The use of codon-usage statistics should improve specificity for CDS.
(iii) Pseudocounts can be added to reflect anticipated sequence error rates in the data examined.
(iv) Since motif matches are not separated into reading frames, frameshifts do not present a problem, except where they occur within motif-matching regions.

Sparse matrices, where individual residues may score zero, are undesirable because the incidence of a codon encoding an amino acid unobserved in the original motif would result in an overall motif match score of zero, even when all the other residue match scores were good. There are two major reasons why this situation can arise: first, the motif may contain relatively few members of the family (most commonly because all the members have yet to be discovered); second, base substituting sequence errors may introduce erroneous codons into the sequence. The first problem can be corrected by the addition of pseudocounts, based on observed substitutions in collections of multiple alignments of related proteins. Such data can be obtained from the BLOSUM (7) or PAM (8) families of substitution matrices. However, because this information is only available at the protein, rather than the DNA, level, the addition of such pseudocounts must be performed prior to cPSSM conversion. The sequence error problem can be compensated for by the addition of codon pseudocounts to the cPSSM. The probabilities of particular codon transitions by sequence errors can be calculated very simply from the anticipated individual base sequence error rate.

Individual and multiple motif *P*-values

In order to better assess the relative strength of motif matches, a system for calculating *P*-values was employed. The goal of an individual cPSSM search is generally to identify the maximal scoring subsequence (MSS) in the search sequence. Once this MSS is identified, the question remains of how significant this match is. Could an MSS as high scoring as the one identified be just as likely to be found in a random or unrelated sequence? Goldstein and Waterman demonstrated that the score distribution of MSSs, $M(n)$, in random sequences approximates to a limiting Gaussian extreme value distribution (GEV) (9). This is based on the assumptions that the subsequence scores at each position in a sequence are independent and that these scores are normally distributed, with mean μ and standard deviation σ. Under the GEV, the probability of getting an MSS in a random sequence with greater than the score observed $m(n)$ can be calculated as follows:

$$P[M(n) \geqslant m(n)] \approx 1 - \exp\left[-\mathrm{e}^{u(n)-m(n)/a(n)}\right], \tag{3}$$

where

$$u(n) = \mu + \sigma\left[\sqrt{2\ln(n)} - \frac{\ln(\ln(n)) + \ln(4\pi)}{2\sqrt{2\ln(n)}}\right] \tag{4}$$

and

$$a(n) = \frac{\sigma}{\sqrt{2\ln(n)}} \tag{5}$$

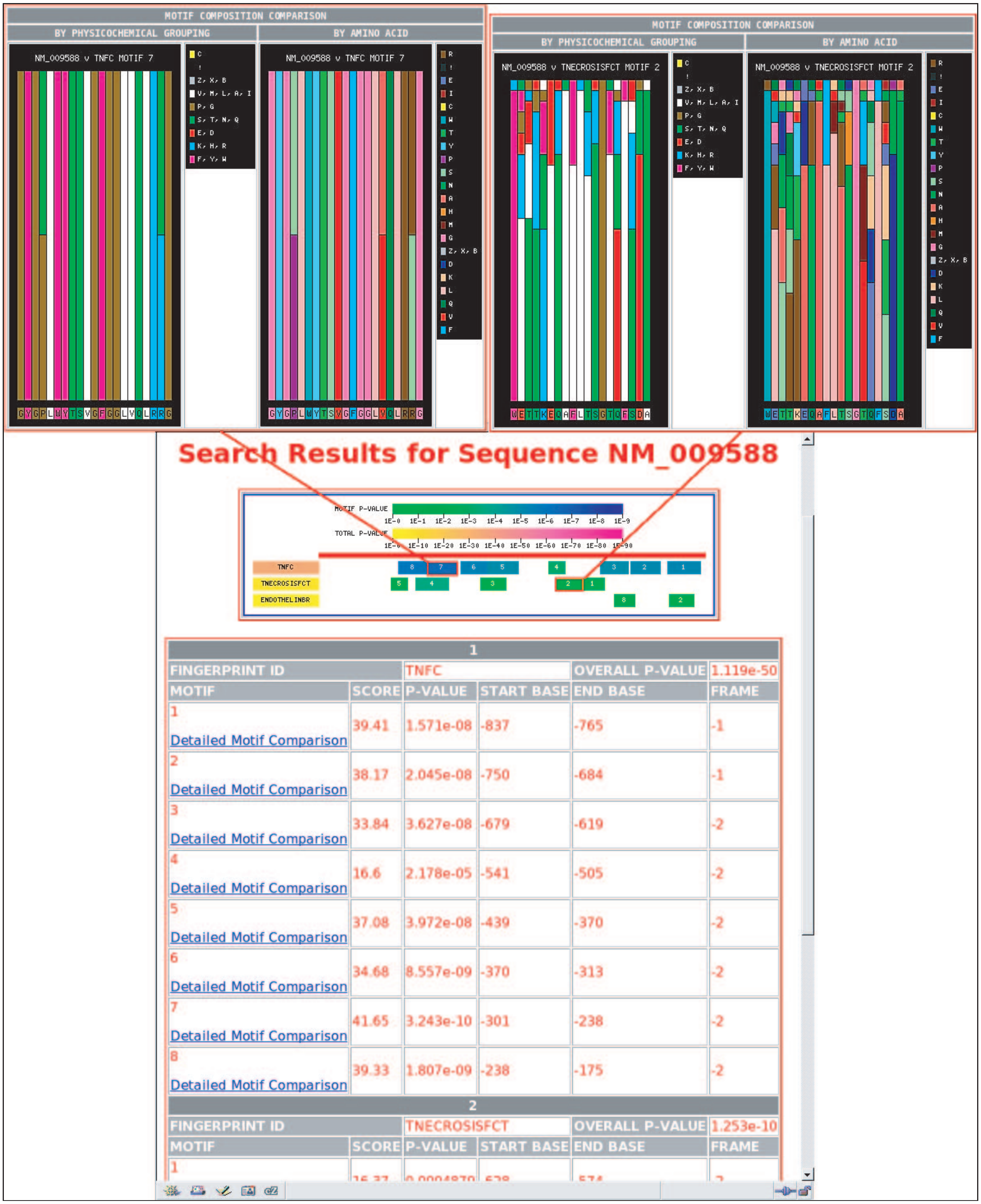
and

$$n = L - W + 1. \tag{6}$$

Furthermore, the average maximum score (or expected value) for a sequence of length L is given by

$$E[M(n)] \approx u(n) + \gamma a(n), \tag{7}$$

where γ is Euler's constant ($\cong 0.5772156649$). Bailey and Gribskov (10) showed that using the observed mean and standard deviations of pseudorandom sequence data to estimate μ and σ results in a poor estimate of $E[M(n)]$ for typical motif widths. However, they further showed that, in such cases, the data appear to be still Gaussian in nature, and better estimates of μ and σ can be obtained by fitting average MSS data to Equation 7. Substituting Equations 4 and 5 into Equation 7, and rearranging into straight-line form, with σ as the gradient


MOTIF COMPOSITION COMPARISON
BY PHYSICOCHEMICAL GROUPING
BY AMINO ACID
NM_009588 v TNFC MOTIF 7
C
!
Z, X, B
V, M, L, A, I
P, G
S, T, N, Q
E, D
K, H, R
F, Y, W
R ! E I C W T Y P S N A H M G Z, X, B D K L Q V F
GYGPLWYTSVGFGGLVQLRRG
MOTIF COMPOSITION COMPARISON
BY PHYSICOCHEMICAL GROUPING
BY AMINO ACID
NM_009588 v TNECROSISFCT MOTIF 2
WETTKEQAFLTSGTQFSDA
MOTIF P-VALUE
1E-0 1E-1 1E-2 1E-3 1E-4 1E-5 1E-6 1E-7 1E-8 1E-9
TOTAL P-VALUE
1E-0 1E-10 1E-20 1E-30 1E-40 1E-50 1E-60 1E-70 1E-80 1E-90
TNFC
8 7 6 5 4 3 2 1
TNECROSISFCT
5 4 3 2 1
ENDOTHELINBR
8 2

Search Results for Sequence NM_009588

1					
FINGERPRINT ID		TNFC		OVERALL P-VALUE	1.119e-50
MOTIF	SCORE	P-VALUE	START BASE	END BASE	FRAME
1 Detailed Motif Comparison	39.41	1.571e-08	-837	-765	-1
2 Detailed Motif Comparison	38.17	2.045e-08	-750	-684	-1
3 Detailed Motif Comparison	33.84	3.627e-08	-679	-619	-2
4 Detailed Motif Comparison	16.6	2.178e-05	-541	-505	-2
5 Detailed Motif Comparison	37.08	3.972e-08	-439	-370	-2
6 Detailed Motif Comparison	34.68	8.557e-09	-370	-313	-2
7 Detailed Motif Comparison	41.65	3.243e-10	-301	-238	-2
8 Detailed Motif Comparison	39.33	1.807e-09	-238	-175	-2
2					
FINGERPRINT ID		TNECROSISFCT		OVERALL P-VALUE	1.253e-10
MOTIF	SCORE	P-VALUE	START BASE	END BASE	FRAME
1	16.37	0.0004870	628	574	2

Figure 1. Screenshot of the results of a query with the mRNA sequence of human tumour necrosis factor C (lymphotoxin-β).

and μ as the intercept, allows the use of linear regression to fit pseudorandom sequence data and estimate better values of μ and σ. Searches on random sequence data suggest that this model can also be used to estimate P-values for cPSSM searches, since adequate goodness-of-fit probabilities are attained for the vast majority of motifs in the PRINTS database (73.5 and 98.8%; $P > 0.001$ and $P > 1 \times 10^{-18}$, respectively).

A WEB INTERFACE

The search algorithm was implemented in a C++ program and resides on a server as a daemon process. A CGI script implemented in perl provides a web interface to the search server. The initial form accepts individual and multiple nucleotide sequence queries up to a maximum total data of 100 Kb (100 000 bases). Search results are presented in tabular form, with an accompanying graphical schematic showing the sequence and significant fingerprint matches, with the locations of all motifs shown relative to their position in the sequence. Both fingerprint and individual motif matches are coloured with a relative hue representing the associated *P*-value. For motifs, colours range from green to blue (weak and strong matches, respectively) and, for whole fingerprints, colours range from yellow to magenta. A screenshot of an example results page is shown in Figure 1.

The query was a human mRNA sequence for a gene encoding tumour necrosis factor C (TNFC) and, as would be expected, the sequence matches the TNFC fingerprint highly significantly ($P = 1.119 \times 10^{-50}$) on the negative strand. There is a second significant match to the TNF family fingerprint (TNECROSISFCT), but the level of significance is much lower because the family-level fingerprint encodes more divergent sequences. The third match, ENDOTHELINBR, achieves a fairly low *P*-value (8.134×10^{-4}), which is similar to an *E*-value of one for a database the size of PRINTS—not surprisingly, the fingerprint is unrelated. The sequence shown in the figure has a frameshift-causing sequence error; however, because it does not occur within a motif-matching region, it has had no effect on the significance of the match. The relative significance of the motif matches is immediately obvious, with the more significant motifs of the TNFC fingerprint standing out in dark blue. In contrast, the individual motif matches to the TNECROSISFCT fingerprint are relatively weak, but all motifs match, and together they give a significant match.

In this sort of situation, where motif matches are weak, it can be helpful to examine how well the translated sequence matches the motifs. However, it can be difficult to visualize the composition of a motif, when, as is often the case for family and superfamily fingerprints, there are many contributing sequences. To overcome this problem, the motifs are represented as histograms, coloured by either amino acid or the physicochemical group to which the amino acid belongs. The histograms for two of the motifs in the example search are shown above the results page in Figure 1. Each bar represents a column in the motif and each bar is subdivided, the relative size of each subdivision being proportional to the contribution of that group or amino acid to the motif column. The histogram for a particular motif can be obtained by clicking on it. A motif from the subfamily (TNFC) and one from the family (TNECROSISFCT) are shown in the figure. The high level of conservation in the subfamily-level motif, which the translated nucleotide subsequence matches almost identically, can be seen at the level of both physicochemical grouping and the individual amino acid. In contrast, at the family level, the motif is much more diverse, but the histogram illustrates that all amino acids observed in the sequence are represented in the appropriate column of the motif, even though they may not always be the most common residue.

DISCUSSION

There is an ever-expanding list of completed and ongoing EST and genome sequencing projects for various species. Online software to facilitate the identification of genes in the data and to infer possible functions is crucial to the various genome annotators. Often pairwise alignment techniques (BLAST, Smith–Waterman, etc.) are used to infer homology and thus function. This approach has proved adequate for the model genetic organisms that were sequenced first. However, as more and more diverse species, which may not be well represented in current sequence databases, are added to the list of genomes to be sequenced, the ability to infer homology between more distant relatives becomes important. Protein family databases contain multiple alignment-based descriptors that contain greater evolutionary information than single protein sequences and can thus help to infer homology between more distantly related sequences. FAN is a server for the PRINTS fingerprint database specifically intended for the submission of nucleotide sequence searches. The use of codon-usage statistics in a cPSSM approach provides a search that is more specific for CDS over non-CDS, and allows the adoption of pseudocount schemes that reflect the anticipated rate of base substitution sequence errors in the data examined (the online version of the software is optimized for an anticipated error rate of 5%). Finally, direct comparison of the descriptor with the nucleotide sequence allows the combination of motif scores in different reading frames, minimizing the effect of frameshift-causing sequence errors.

ACKNOWLEDGEMENTS

We are grateful to the EPSRC for individual support (N.M.).

REFERENCES

1. Das,M., Burge,C., Park,E., Colinas,J. and Pelletier,J. (2001) Assessment of the total number of human transcription units. *Genomics*, **77**, 71–78.
2. Guigo,R., Agarwal,P., Abril,J., Burset,M. and Fickett,J. (2000) An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.*, **10**, 1631–1642.
3. Attwood,T., Bradley,P., Flower,D., Gaulton,A., Maudling,N., Mitchell,A., Moulton,G., Nordle,A., Paine,K., Taylor,P., Uddin,A. and Zygouri,C. (2003) PRINTS and its automatic supplement, preprints. *Nucleic Acids Res.*, **31**, 400–402.
4. Henikoff,J., Greene,E., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
5. Claverie,J. (1997) Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.*, **6**, 1735–1744.
6. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*, Cambridge University Press, Cambridge, UK.
7. Henikoff,S. and Henikoff,J. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
8. Dayhoff,M., Schwartz,R. and Orcutt,B. (1978) A model of evolutionary change in proteins. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Washington. Vol. 5, suppl. 3, pp. 345–352.
9. Goldstein,L. and Waterman,M. (1994) Approximations to profile score distributions. *J. Comput. Biol.*, **1**, 93–104.
10. Bailey,T. and Gribskov,M. (1997) Score distributions for simultaneous matching to multiple motifs. *J. Comput. Biol.*, **4**, 45–59.